Statistical Physics of Information Measures

Neri Merhav

Department of Electrical Engineering

Technion – Israel Institute of Technology

Technion City, Haifa, Israel

Partly joint work with **D. Guo** (Northwestern U.) and **S. Shamai** (Technion).

Physics of Algorithms '09, Santa Fe, NM, USA, Aug. 31 – Sep. 4, 2009

Outline

Relations between Information Theory (IT) and statistical physics:

- Conceptual aspects – relations between principles in the two areas.
- Technical aspects – identifying similar mathematical formalisms and borrowing techniques.

In this talk we:

- Briefly review basic background in IT.
- Discuss some physics of the Shannon limits.
- Briefly review basic background in estimation theory.
- Touch upon statistical physics of signal estimation via the mutual information.

First Part:

Physics of the Shannon Limits

The Shannon Limits

- Lossless data compression:

  compression ratio $\geq H =$ entropy.

- Lossy compression:

  compression ratio $\geq R(D) =$ rate–distortion func.

- Channel coding:

  coding rate $\leq C =$ channel capacity.

- Joint source–channel coding:

  decoding error $\geq R^{-1}(C) =$ distortion–rate func. at rate C.

- etc. etc. etc.

The Information Inequality

Each of the above–mentioned fundamental limits of IT, as well as many others, is based on the information inequality in some form:

For any two distributions, P and Q, over an alphabet $\mathcal{X}$:

$$D(P\|Q) \stackrel{\Delta}{=} \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \geq 0.$$

In physics, it is known as the Gibbs inequality.

The Gibbs Inequality

Let $\mathcal{E}_0(x)$ and $\mathcal{E}_1(x)$ be two Hamiltonians of a system. For a given β, let

$$P_i(x) = \frac{e^{-\beta \mathcal{E}_i(x)}}{Z_i}, \qquad Z_i = \sum_x e^{-\beta \mathcal{E}_i(x)}, \quad i = 0, 1.$$

Then,

$$\begin{aligned} 0 \quad &\le \quad D(P_0 \| P_1) = \left\langle \ln \frac{e^{-\beta \mathcal{E}_0(X)}/Z_0}{e^{-\beta \mathcal{E}_1(X)}/Z_1} \right\rangle_0 \\ &= \quad \ln Z_1 - \ln Z_0 + \beta \left\langle \mathcal{E}_1(X) - \mathcal{E}_0(X) \right\rangle_0 \end{aligned}$$

or

$$\begin{aligned} \left\langle \mathcal{E}_1(X) - \mathcal{E}_0(X) \right\rangle_0 \quad &\ge \quad kT \ln Z_0 - kT \ln Z_1 \\ &= \quad F_1 - F_0 \end{aligned}$$

Interpretation of $\langle \mathcal{E}_1(X) - \mathcal{E}_0(X) \rangle_0 \geq \Delta F$

- A system with Hamiltonian $\mathcal{E}_0(x)$ – in equilibrium $\forall\ t < 0$.
  Free energy $= -kT \ln Z_0$.
- At $t = 0$, the Hamiltonian jumps, by $W = \mathcal{E}_1(x) - \mathcal{E}_0(x)$: from $\mathcal{E}_0(x)$ to $\mathcal{E}_1(x)$ – by abruptly applying a force. Energy injected:
  $\langle W \rangle_0 = \langle \mathcal{E}_1(X) - \mathcal{E}_0(X) \rangle_0$.
- New system, with Hamiltonian $\mathcal{E}_1$, equilibrates.
  Free energy $= -kT \ln Z_1$.

Gibbs inequality: $\langle W \rangle_0 \geq \Delta F$.

$$\langle W \rangle_0 - \Delta F = kT \cdot D(P_0 \| P_1)$$

is the dissipated energy $=$ entropy production (system $+$ environment) due to irreversibility of the abruptly applied force.

Example – Data Compression and the Ising Model

Let $\boldsymbol{X} \in \{-1,+1\}^n \ \sim$ Markov chain $P_0(\boldsymbol{x}) = \prod_i P_0(x_i|x_{i-1})$ with

$$P_0(x|x') = \frac{\exp(Jx \cdot x')}{Z_0}, \qquad x, x' \in \{-1,+1\}$$

Code designer thinks that $\boldsymbol{X} \ \sim$ Markov with parameters:

$$P_1(x|x') = \frac{\exp(Jx \cdot x' + Kx)}{Z_1(x')}.$$

$D(P_0\|P_1) =$ loss in compression due to mismatch. Easy to see that

$$\mathcal{E}_0(\boldsymbol{x}) = -J \cdot \sum_i x_i x_{i-1}; \quad \mathcal{E}_1(\boldsymbol{x}) = -J \cdot \sum_i x_i x_{i-1} - B \cdot \sum_i x_i$$

where

$$B = K + \frac{1}{2} \ln \frac{\cosh(J-K)}{\cosh(J+K)}.$$

Thus, $W = -B \cdot \sum_i x_i$ means an abrupt application of the magnetic field B.

Physics of the Data Processing Theorem (DPT)

Mutual information: Let $(U,V) \sim P(u,v)$:

$$I(U;V) \equiv \left\langle \log \frac{P(U,V)}{P(U)P(V)} \right\rangle .$$

DPT:

$$X \to U \to V \text{ Markov chain } \implies I(X;U) \geq I(X;V).$$

Pf:

$$I(X;U) - I(X;V) = \left\langle D(P_{X|U,V}(\cdot|U,V) \| P_{X|V}(\cdot|V)) \right\rangle \geq 0. \quad \Box$$

Supports most, if not ∀, Shannon limits.

Physics of the DPT (Cont'd)

Let $\beta = 1$. Given (u, v), let

$$\mathcal{E}_0(x) = -\ln P(x|u, v) = -\ln P(x|u); \quad \mathcal{E}_1(x) = -\ln P(x|v).$$

$$Z_0 = \sum_x e^{-1\cdot[-\ln P(x|u,v)]} = \sum_x P(x|u, v) = 1$$

and similarly, $Z_1 = 1$. Thus, $F_0 = F_1 = 0$, and so, $\Delta F = 0$.
After averaging over P_{UV}:

$$\begin{aligned} \langle W(X) \rangle_0 &= \langle -\ln P(X|V) + \ln P(X|U) \rangle \\ &= H(X|V) - H(X|U) \\ &= I(X;U) - I(X;V). \end{aligned}$$

$$\langle W \rangle_0 = I(X;U) - I(X;V) \geq 0 = \Delta F.$$

Discussion

The relation

$$\langle W\rangle_0 - \Delta F = kT\cdot D(P_0\|P_1)\geq 0$$

is known (Jarzynski ‘97, Crooks ‘99, ..., Kawai *et. al.* ‘07), but with different physical interpretations, which require some limitations.

Present interpretation – holds generally; Applied in particular to the DPT.

In our case:

- Maximum irreversibility: $\langle W\rangle_0$ – fully dissipated: $\Delta F = 0$.
- All dissipation – in the system, none in the environment:

$$\langle W\rangle_0 = T\Delta S = 1\cdot[H(X|V) - H(X|U)].$$

- Rate loss due to gap between mutual informations: irreversible process $\Longleftrightarrow$ irreversible info: $I(X;U) > I(X;V) \longrightarrow U$ cannot be retrieved from V.

Relation to Jarzynski's Equality

Let

$$\mathcal{E}_\lambda(x) = \mathcal{E}_0(x) + \lambda[\mathcal{E}_1(x) - \mathcal{E}_0(x)]$$

interpolate $\mathcal{E}_0$ and $\mathcal{E}_1$. λ – a generalized force.
Jarzynski's equality (1997): $\forall$ protocol $\{\lambda_t\}$ with $\lambda_t = 0 \; \forall \; t \leq 0$ and $\lambda_t = 1$ $\forall \; t \geq \tau$ ($\tau \geq 0$), the injected energy

$$W = \int_0^\tau \mathrm{d}\lambda_t [\mathcal{E}_1(x_t) - \mathcal{E}_0(x_t)]$$

satisfies

$$\left\langle e^{-\beta W} \right\rangle = e^{-\beta \Delta F}.$$

Jensen: $\left\langle e^{-\beta W} \right\rangle \geq \exp\{-\beta \langle W \rangle\}$ so, $\langle W \rangle \geq \Delta F$ more generally.
Equality – for a reversible process – $W =$ deterministic.

Informational Jarzynski Equality

Taking

$$\mathcal{E}_0(x) = -\ln P_0(x), \quad \mathcal{E}_1(x) = -\ln P_1(x), \quad \beta = 1$$

and defining a “protocol” $0 \equiv \lambda_0 \to \lambda_1 \to \ldots \to \lambda_n \equiv 1$, and

$$W = \sum_{i=0}^{n-1} (\lambda_{i+1} - \lambda_i) \ln \frac{P_0(X_i)}{P_1(X_i)}, \quad X_i \sim P_{\lambda_i} \propto P_0^{1-\lambda_i} P_1^{\lambda_i},$$

one can show:

$$\left\langle e^{-W} \right\rangle = 1 = e^{-\Delta F}.$$

Jensen: generalized information inequality:

$$\int_0^1 \mathrm{d}\lambda_t \left\langle \ln \frac{P_0(X)}{P_1(X)} \right\rangle_{\lambda_t} \geq 0.$$

Summary of First Part

- Suboptimum commun. system $\Longleftrightarrow$ irreversible process.
- Info rate loss $\Longleftrightarrow$ dissipated energy $\rightarrow$ entropy $\uparrow$
- Fundamental limits of IT $\Longleftrightarrow$ second law.
- Possible implications of Jarzynski's equality in IT.

Second Part:

Statistical Physics of Signal Estimation via the Mutual Information

Signal Estimation – Preliminaries

Let

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z} \qquad \text{(all vectors in } \mathbb{R}^n\text{)}$$

where $\boldsymbol{X}$ is the desired signal and $\boldsymbol{Z}$ is noise $\perp \boldsymbol{X}$.
Estimator: any function $\hat{\boldsymbol{X}} = f(\boldsymbol{Y})$. We want $\hat{\boldsymbol{X}}$ as 'close' as possible to $\boldsymbol{X}$.

$$\text{mean square error} = \left\langle \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \right\rangle = \left\langle \|\boldsymbol{X} - f(\boldsymbol{Y})\|^2 \right\rangle.$$

A fundamental result: minimum mean square error (MMSE) $=$ conditional mean:

$$\boldsymbol{X}^* = f^*(\boldsymbol{y}) = \langle \boldsymbol{X} \rangle_{\boldsymbol{Y}=\boldsymbol{y}} \equiv \int \mathrm{d}\boldsymbol{x} \cdot \boldsymbol{x} P(\boldsymbol{x}|\boldsymbol{y}).$$

Normally – difficult to apply $\boldsymbol{X}^*$ and assess performance.

$\boldsymbol{X}^*$ and MMSE may exhibit irregularities – threshold effects $\longleftrightarrow$ phase transitions in analogous physical systems. Motivates a statistical–mechanical perspective.

The I–MMSE Relation

[Guo–Shamai–Verdú 2005]: for $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}$, $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I} \cdot 1/\beta)$, regardless of $P(\boldsymbol{X})$:

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = 2 \cdot \frac{\text{d}}{\text{d}\beta} I(\boldsymbol{X}; \boldsymbol{Y}),$$

where $\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) \equiv \langle \|\boldsymbol{X} - f^*(\boldsymbol{Y})\|^2 \rangle$.

Simple example: If $\boldsymbol{X} \sim \mathcal{N}(0, \sigma^2 I)$,

$$\frac{I(\boldsymbol{X}; \boldsymbol{Y})}{n} = \frac{1}{2} \log(1 + \beta\sigma^2)$$

$$\Longrightarrow \quad \frac{\text{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = \frac{\sigma^2}{1 + \beta\sigma^2}.$$

MMSE – calculated using stat–mech via the mutual info and I–MMSE relation $\Longrightarrow$

Statistical Physics of the MMSE

$$\begin{aligned} I(\boldsymbol{X};\boldsymbol{Y}) &= \left\langle \log \frac{P(\boldsymbol{X}|\boldsymbol{Y})}{P(\boldsymbol{X})} \right\rangle_\beta \\ &= \left\langle \log \frac{\exp\{-\beta\|\boldsymbol{Y}-\boldsymbol{X}\|^2/2\}}{\sum_{\boldsymbol{x}} P(\boldsymbol{x})\exp\{-\beta\|\boldsymbol{Y}-\boldsymbol{x}\|^2/2\}} \right\rangle_\beta \\ &= -\frac{n}{2} - \langle \log Z(\beta|\boldsymbol{Y})\rangle_\beta \end{aligned}$$

where

$$Z(\beta|\boldsymbol{Y}) = \sum_{\boldsymbol{x}} P(\boldsymbol{x})\exp\{-\beta\|\boldsymbol{Y}-\boldsymbol{x}\|^2/2\},$$

and so,

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = 2\cdot\frac{\mathrm{d}I(\boldsymbol{X};\boldsymbol{Y})}{\mathrm{d}\beta} = -2\frac{\partial}{\partial\beta}\langle \log Z(\beta|\boldsymbol{Y})\rangle_\beta.$$

Similar to internal energy, but here also $\langle\cdot\rangle_\beta$ depends on β.

Statistical Physics of the MMSE (Cont’d)

A more detailed derivation yields:

$$\text{mmse}(\boldsymbol{X}|\boldsymbol{Y}) = \frac{n}{\beta} + \text{Cov}\{\|\boldsymbol{Y} - \boldsymbol{X}\|^2, \log Z(\beta|\boldsymbol{Y})\}$$

- The term $n/\beta \sim$ energy equipartition theorem.
- Covariance term – dependence of $\langle\cdot\rangle_\beta$ on β.

Statistical Physics of the MMSE (Cont'd)

In stat. mech: $$\begin{aligned}\Sigma(\beta) &= \log Z(\beta) + \beta\langle\mathcal{E}(X)\rangle \\ &= \log Z(\beta) - \beta\frac{\text{d}\log Z(\beta)}{\text{d}\beta} \quad \Longleftarrow \text{ diff. eq.}\end{aligned}$$

$$\log Z(\beta) = -\beta E_0 + \beta \cdot \int_\beta^\infty \frac{\text{d}\hat{\beta}\cdot\Sigma(\hat{\beta})}{\hat{\beta}^2}; \qquad E_0 = \text{ground–state energy}$$

$$\Longrightarrow\; E = -\frac{\text{d}\log Z(\beta)}{\text{d}\beta} = \left[E_0 - \int_\beta^\infty \frac{\text{d}\hat{\beta}\cdot\Sigma(\hat{\beta})}{\hat{\beta}^2}\right] + \frac{\Sigma(\beta)}{\beta}$$

Similarly for $\langle\log Z(\beta|\boldsymbol{Y})\rangle_\beta$ except that

$$\Sigma(\beta) \Longleftarrow \frac{\beta}{2}\text{Cov}\{\|\boldsymbol{Y}-\boldsymbol{X}\|^2, \log Z(\beta|\boldsymbol{Y})\} - I(\boldsymbol{X};\boldsymbol{Y})$$

$$E_0 \Longleftarrow \frac{1}{2}\left\langle\min_{\boldsymbol{x}}\|\boldsymbol{Y}-\boldsymbol{x}\|^2\right\rangle_\beta .$$

Examples

Example 1 – Random Codebook on a Sphere Surface

$$\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{Z}; \qquad \boldsymbol{X} \sim \text{Unif}\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}, \;\; M = e^{nR}$$

Codewords: randomly drawn independently uniformly on Surf$(\sqrt{n\sigma^2})$.

$$\lim_{n\to\infty} \frac{\langle I(\boldsymbol{X};\boldsymbol{Y})\rangle}{n} = \begin{cases} \frac{1}{2}\log(1+\beta\sigma^2) & \beta < \beta_R \\ R & \beta \geq \beta_R \end{cases}$$

where β_R is the solution to the eqn $R = \frac{1}{2}\log(1+\beta\sigma^2)$. Thus,

$$\lim_{n\to\infty} \frac{\text{mmse}(\boldsymbol{X}|\boldsymbol{Y})}{n} = \begin{cases} \frac{\sigma^2}{1+\beta\sigma^2} & \beta < \beta_R \\ 0 & \beta \geq \beta_R \end{cases}$$

A 1st–order ϕ transition in MMSE: At high temp. behaves as if $\boldsymbol{X}$ was Gaussian and at $\beta = \beta_R$ jumps to zero!

Examples (Cont’d)

Example 2 – Sparse Signals

$$X_i = \left(\frac{1-\mu_i}{2}\right) U_i, \quad i = 1, \ldots, n$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n) \sim P(\boldsymbol{\mu})$ are binary $\{\pm 1\}$; $U_i \sim \mathcal{N}(0, \sigma^2)$ – i.i.d. $\perp$ $\boldsymbol{\mu}$.

$$\begin{aligned}
Z(\beta|\boldsymbol{y}) &= \int_{\mathbb{R}^n} \mathrm{d}\boldsymbol{x} P(\boldsymbol{x}) \exp\{-\beta \|\boldsymbol{y}-\boldsymbol{x}\|^2/2\} \quad \Longleftarrow \quad P(\boldsymbol{x}) = \sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) P(\boldsymbol{x}|\boldsymbol{\mu}) \\
&= \sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} \text{func}(y_i, \mu_i, q)\right\} \quad \Longleftarrow \quad q \equiv \beta\sigma^2 \\
&= \text{const.} \times \sum_{\boldsymbol{\mu}} P(\boldsymbol{\mu}) \cdot \exp\left\{\sum_{i=1}^{n} \mu_i h_i\right\} \quad h_i = \text{func}(y_i)
\end{aligned}$$

Sum over $\{\boldsymbol{\mu}\} \equiv \hat{Z}(\beta|\boldsymbol{y})$: “partition function” of spins in a random field $\{h_i\}$.

Example 2 (Cont'd)

Let $P(\boldsymbol{\mu}) \propto \exp\{nf[m(\boldsymbol{\mu})]\}$ where $m(\boldsymbol{\mu}) \equiv \frac{1}{n}\sum_i \mu_i$ and $f[m]$ is 'nice'.

$$\hat{Z}(\beta|\boldsymbol{y}) \propto \sum_{\boldsymbol{\mu}} \exp\left\{n\left[f[m(\boldsymbol{\mu})] + \frac{1}{n}\sum_i \mu_i h_i\right]\right\}$$

$\hat{Z}$ is dominated by configurations with magnetization m^*, solving the zero–derivative equation

$$m = \langle \tanh(f'[m] + H)\rangle$$

where H is a RV pertaining to h_i. $m^* =$ local maximum if:

$$\left\langle \tanh^2(f'[m^*] + H)\right\rangle > 1 - \frac{1}{f''[m^*]}.$$

When this becomes equality (and then reversed), m^* ceases to dominate $\hat{Z}$ (critical point) $\Longrightarrow$ dominant magentization jumps elsewhere.

Example 2 (Cont'd)

Consider the case

$$f[m] = am + \frac{bm^2}{2}$$

$\hat{Z}$ – similar to the random–field Curie–Weiss (RFCW) model.

We analyze the mutual info using stat–mech methods, and then derive the MMSE using the I–MMSE relation:

MMSE for Example 2

$$\begin{aligned}\overline{\text{mmse}} \quad = \quad & \frac{\sigma^2 q}{2(1+q)^2} + \frac{(1-m_a)\sigma^2}{2}\left[1 - \frac{q(1+q/2)}{(1+q)^2}\right] + \\ & \frac{1+m_a}{2}\Big[\mathsf{Cov}_0\{Y^2, \log[2\cosh(bm^* + a + H)]\} + \\ & \left\langle H' \tanh(bm^* + a + H)\right\rangle_0\Big] + \\ & + \frac{1-m_a}{2}\Big[\frac{1}{(1+q)^2} \cdot \mathsf{Cov}_1\{Y^2, \log[2\cosh(bm^* + a + H)]\} + \\ & \left\langle H' \tanh(bm^* + a + H)\right\rangle_1\Big]\end{aligned}$$

where $\langle\cdot\rangle_s$ and Cov$_s$ are w.r.t. $Y \sim \mathcal{N}(0, \sigma^2 s + 1/\beta)$, $s = 0, 1$, and

$$H' = -\frac{\sigma^2}{2(1+q)} + \frac{q(q+2)Y^2}{2(1+q)^2}.$$

Example 2: Discussion

- MMSE depends on m^*: jumps of m^* yield discontinuities in MMSE.
- As m^* jumps, the response of $\boldsymbol{X}^*(\boldsymbol{Y})$ jumps as well.
- In the C–W model: 1st order transition w.r.t. mag. field and 2nd order transition w.r.t. β. Here – a 1st order transition w.r.t. β because dependence on β is via the “magnetic fields” $\{h_i\}$..
- $b = 0$: i.i.d. spins $\Longrightarrow$ no ϕ transitions $\Longrightarrow$ sparsity alone does not cause ϕ transitions.

Conclusion of Second Part

- MMSE calculated using stat. mech. via the mutual info.
- Statistical–mech techniques can be used to inspect inherent irregularities in the estimation error, via phase transitions.
- Possible to handle situations of mismatch between true prior P and assumed prior Q.